Decomposable Graphical Gaussian Models

Steffen Lauritzen, University of Oxford

CIMPA Summerschool, Hammamet 2011, Tunisia

September 12, 2011

This simple algorithm has complexity $O(|V| + |E|)$:

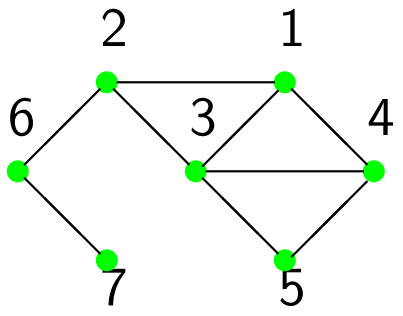
1. Choose $v_0 \in V$ arbitrary and let $v_0 = 1$;
2. When vertices $\{1, 2, \ldots, j\}$ have been identified, choose $v = j + 1$ among $V \setminus \{1, 2, \ldots, j\}$ with highest cardinality of its numbered neighbours;
3. *If* $\mathrm{bd}(j+1) \cap \{1, 2, \ldots, j\}$ *is not complete,* $\mathcal{G}$ is not chordal;
4. Repeat from 2;
5. *If the algorithm continues until no vertices are left, the graph is chordal and the numbering is perfect.*

Finding the cliques of a chordal graph

From an MCS numbering $V = \{1, \ldots, |V|\}$, let

$$B_\lambda = \mathrm{bd}(\lambda) \cap \{1, \ldots, \lambda - 1\}$$

and $\pi_\lambda = |B_\lambda|$. Call λ a *ladder vertex* if $\lambda = |V|$ or if $\pi_{\lambda+1} < \pi_\lambda + 1$. Let Λ be the set of ladder vertices.



π_λ: 0,1,2,2,2,1,1.

The cliques are $C_\lambda = \{\lambda\} \cup B_\lambda, \lambda \in \Lambda$.

Let $\mathcal{A}$ be a collection of finite subsets of a set V. A *junction tree* $\mathcal{T}$ of sets in $\mathcal{A}$ is an undirected tree with $\mathcal{A}$ as a vertex set, satisfying the *junction tree property:*

> *If $A, B \in \mathcal{A}$ and C is on the unique path in $\mathcal{T}$ between A and B it holds that $A \cap B \subset C$.*

If the sets in an arbitrary $\mathcal{A}$ are pairwise incomparable, *they can be arranged in a junction tree if and only if $\mathcal{A} = \mathcal{C}$ where $\mathcal{C}$ are the cliques of a chordal graph*

The following are equivalent for any undirected graph $\mathcal{G}$.

(i) *$\mathcal{G}$ is chordal;*

(ii) *$\mathcal{G}$ is decomposable;*

(iii) *All prime components of $\mathcal{G}$ are cliques;*

(iv) *$\mathcal{G}$ admits a perfect numbering;*

(v) *Every minimal (α, β)-separator are complete.*

(vi) *The cliques of $\mathcal{G}$ can be arranged in a junction tree.*

The junction tree can be *constructed directly from the MCS ordering* $C_\lambda, \lambda \in \Lambda$, where C_λ are the cliques: Since the MCS-numbering is perfect, $C_\lambda, \lambda > \lambda_{\min}$ all satisfy

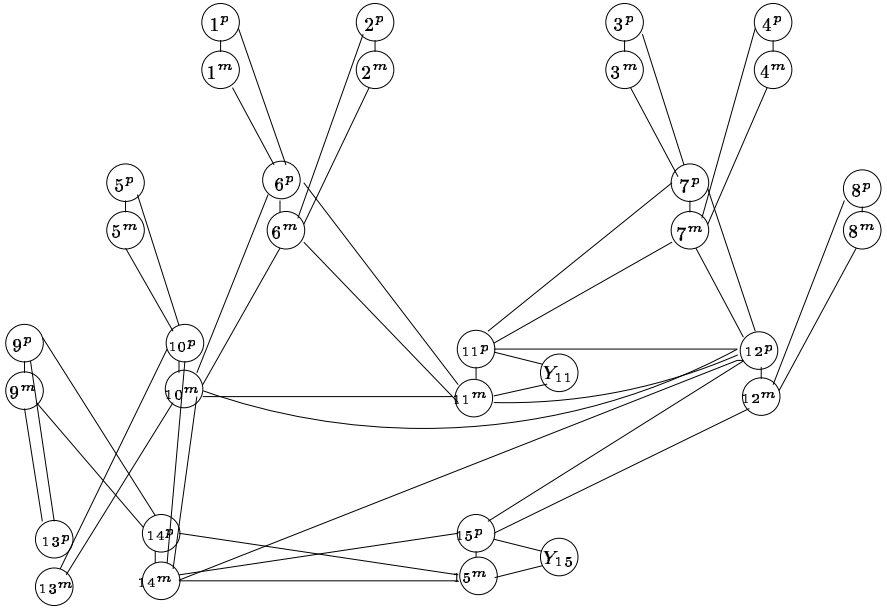
$$C_\lambda \cap (\cup_{\lambda' < \lambda} C_{\lambda'}) = C_\lambda \cap C_{\lambda^*} = S_\lambda$$

for some $\lambda^* < \lambda$.

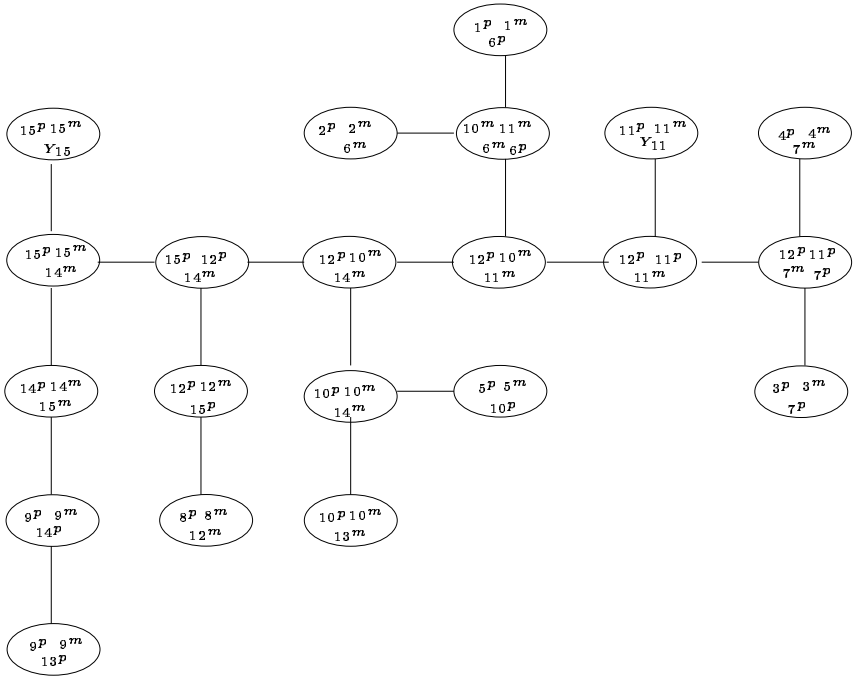
A junction tree is now easily constructed by attaching C_λ to any C_{λ^*} satisfying the above. Although λ^* may not be uniquely determined, S_λ is.

Indeed, the sets S_λ are the minimal complete separators and *the numbers* $\nu(S)$ *are* $\nu(S) = |\{\lambda \in \Lambda : S_\lambda = S\}|$.

A chordal graph



Junction tree



Cliques of graph arranged into a tree with $C_1 \cap C_2 \subseteq D$ for all cliques D on path between C_1 and C_2.

In general, the *prime components* of any undirected graph can be arranged in a junction tree in a similar way.

Then *every pair of neighbours* (C, D) *in the junction tree represents a decomposition of* $\mathcal{G}$ into $\mathcal{G}_{\tilde{C}}$ and $\mathcal{G}_{\tilde{D}}$, where $\tilde{C}$ is the set of vertices in prime components connected to C but separated from D in the junction tree, and similarly with $\tilde{D}$.

The corresponding algorithm is based on a slightly more sophisticated algorithm known as *Lexicographic Search* (LEX) which runs in $O(|V|^2)$ time.

If the graph $\mathcal{G}$ is chordal, we say that the graphical model is *decomposable*.

In this case, *the IPS-algorithm converges in a finite number of steps*.

We also have the familiar *factorization of densities*

$$f(x \mid \Sigma) = \frac{\prod_{C \in \mathcal{C}} f(x_C \mid \Sigma_C)}{\prod_{S \in \mathcal{S}} f(x_S \mid \Sigma_S)^{\nu(S)}} \tag{1}$$

where $\nu(S)$ is the number of times S appear as intersection between neighbouring cliques of a junction tree for $\mathcal{C}$.

Relations for trace and determinant

Using the factorization (1) we can for example match the expressions for the trace and determinant of Σ

$$\mathrm{tr}(KW) = \sum_{C \in \mathcal{C}} \mathrm{tr}(K_C W_C) - \sum_{S \in \mathcal{S}} \nu(S)\, \mathrm{tr}(K_S W_S)$$

and further

$$\det \Sigma \quad = \quad \{\det(K)\}^{-1} = \frac{\prod_{C \in \mathcal{C}} \det\{\Sigma_C\}}{\prod_{S \in \mathcal{S}} \{\det(\Sigma_S)\}^{\nu(S)}}$$

These are some of many relations that can be derived using the decomposition property of chordal graphs.

The same factorization clearly holds for the maximum likelihood estimates:

$$f(x \mid \hat{\Sigma}) = \frac{\prod_{C \in \mathcal{C}} f(x_C \mid \hat{\Sigma}_C)}{\prod_{S \in \mathcal{S}} f(x_S \mid \hat{\Sigma}_S)^{\nu(S)}} \tag{2}$$

Moreover, it follows from the general likelihood equations that

$$\hat{\Sigma}_A = W_A/n \text{ whenever } A \text{ is complete.}$$

Exploiting this, we can obtain an explicit formula for the maximum likelihood estimate in the case of a chordal graph.

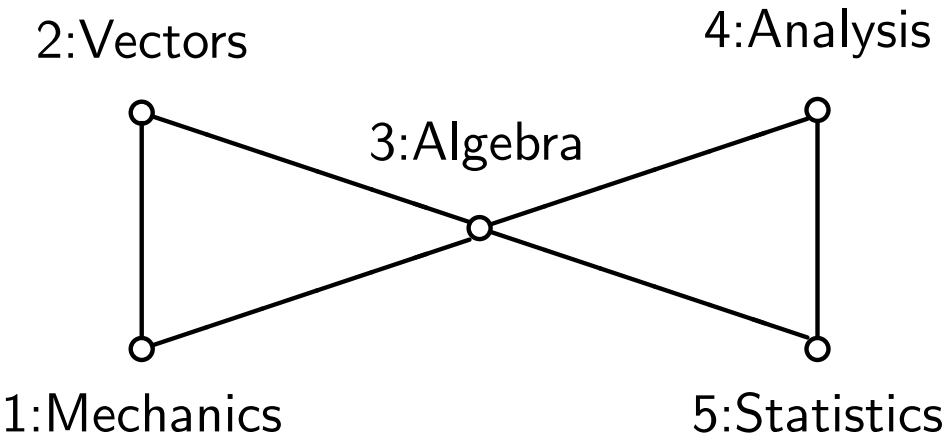
For a $|d| \times |e|$ matrix $A = \{a_{\gamma\mu}\}_{\gamma \in d, \mu \in e}$ we let $[A]^V$ denote the matrix obtained from A by filling up with zero entries to obtain full dimension $|V| \times |V|$, i.e.

$$\left([A]^V\right)_{\gamma\mu} = \begin{cases} a_{\gamma\mu} & \text{if } \gamma \in d, \mu \in e \\ 0 & \text{otherwise.} \end{cases}$$

The maximum likelihood estimates exists if and only if $n \geq C$ for all $C \in \mathcal{C}$. Then the following simple formula holds for the maximum likelihood estimate of K:

$$\hat{K} = n \left\{ \sum_{C \in \mathcal{C}} \left[(w_C)^{-1} \right]^V - \sum_{S \in \mathcal{S}} \nu(S) \left[(w_S)^{-1} \right]^V \right\}.$$

Mathematics marks



This graph is chordal with cliques $\{1, 2, 3\}$, $\{3, 4, 5\}$ with separator $S = \{3\}$ having $\nu(\{3\}) = 1$.

Since one degree of freedom is lost by subtracting the average, we get in this example

$$\hat{K} = 87 \begin{pmatrix} w^{11}_{[123]} & w^{12}_{[123]} & w^{13}_{[123]} & 0 & 0 \\ w^{21}_{[123]} & w^{22}_{[123]} & w^{23}_{[123]} & 0 & 0 \\ w^{31}_{[123]} & w^{32}_{[123]} & w^{33}_{[123]} + w^{33}_{[345]} - 1/w_{33} & w^{34}_{[345]} & w^{35}_{[345]} \\ 0 & 0 & w^{43}_{[345]} & w^{44}_{[345]} & w^{45}_{[345]} \\ 0 & 0 & w^{53}_{[345]} & w^{54}_{[345]} & w^{55}_{[345]} \end{pmatrix}$$

where $w^{ij}_{[123]}$ is the ijth element of the inverse of

$$W_{[123]} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{pmatrix}$$

and so on.

The formula

$$\hat{\Sigma} = n^{-1} \left\{ \sum_{C \in \mathcal{C}} \left[(W_C)^{-1} \right]^V - \sum_{S \in \mathcal{S}} \nu(S) \left[(W_S)^{-1} \right]^V \right\}^{-1}$$

specifies $\hat{\Sigma}$ as a random matrix.

The distribution of this random Wishart-type matrix is partly reflecting Markov properties of the graph $\mathcal{G}$.

This is also true for the distribution of $\hat{\Sigma}$ for a non-chordal graph $\mathcal{G}$ but not to the same degree.

Before we delve further into this, we shall need some more terminology.

Laws and distributions

Generally we think of

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}$$

and sometimes identify Θ with $\mathcal{P}$ which is justified when the parametrization

$$\theta \to P_\theta$$

is one-to-one and onto.

In a Gaussian graphical model $\theta = K \in \mathcal{S}^+(\mathcal{G})$ is uniquely identifying any regular Gaussian distribution satisfying the Markov properties w.r.t. $\mathcal{G}$.

In any case, any probability measure on $\mathcal{P}$ (or on Θ) represents a random element of $\mathcal{P}$, i.e. a random distribution. The *sampling distribution of the MLE* $\hat{\Sigma}$ is an example of such a measure.

To keep heads straight we refer to a probability measure on $\mathcal{P}$ as a *law*, whereas a *distribution* is a probability measure on $\mathcal{X}$.

Thus we shall e.g. speak of the *Wishart law* as we think of it specifying a distribution of $f(\cdot \mid \Sigma)$.

We identify $\theta \in \Theta$ and $P_\theta \in \mathcal{P}$, so e.g. θ_A for $A \subseteq V$ denotes the distribution of X_A under P_θ and $\theta_{A \mid B}$ the family of conditional distributions of X_A given X_B, etc.

For a law $\mathcal{L}$ on Θ we write

$$A \perp\!\!\!\perp_{\mathcal{L}} B \mid S \iff \theta_{A \cup S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \cup S} \mid \theta_S.$$

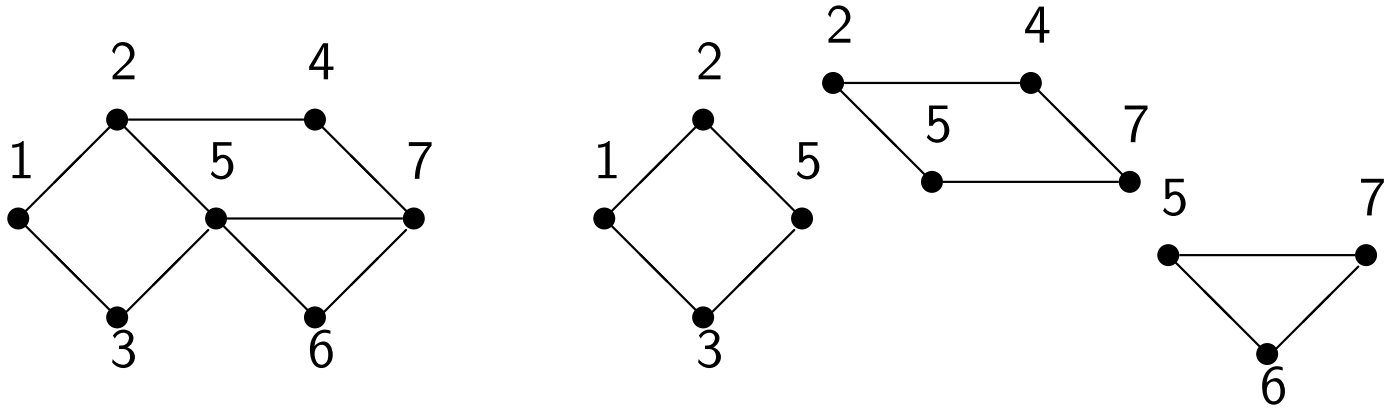
A law $\mathcal{L}$ on Θ is *hyper Markov* w.r.t. $\mathcal{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathcal{G}$;

(ii) $A \perp\!\!\!\perp_{\mathcal{L}} B \mid S$ whenever S is complete and $A \perp_{\mathcal{G}} B \mid S$.

Note the conditional independence is only required to hold for *graph decompositions.*

Hyper Markov property



If θ follows a hyper Markov law for this graph, it holds for example that

$$\theta_{1235} \perp\!\!\!\perp \theta_{24567} \mid \theta_{25}.$$

This is indeed true for $\hat{\Sigma}$ in the graphical model with this graph, i.e.

$$\hat{\Sigma}_{1235} \perp\!\!\!\perp \hat{\Sigma}_{24567} \mid \hat{\Sigma}_{25}.$$

Consequences of the hyper Markov property

Clearly, if $A \perp\!\!\!\perp_{\mathcal{L}} B \mid S$, we have for example also (using property (C2) of conditional independence)

$$\theta_A \perp\!\!\!\perp_{\mathcal{L}} \theta_B \mid \theta_S$$

since θ_A and θ_B are functions of $\theta_{A \cup S}$ and $\theta_{B \cup S}$ respectively.

But *the converse is false!* $\theta_A \perp\!\!\!\perp_{\mathcal{L}} \theta_B \mid \theta_S$ does *not* imply $\theta_{A \cup S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \cup S} \mid \theta_S$, since $\theta_{A \cup S}$ is *not* a function of (θ_A, θ_S). In contrast, $X_{A \cup B}$ is indeed a (one-to-one) function of (X_A, X_B).

However *it generally holds* that

$$A \perp\!\!\!\perp_{\mathcal{L}} B \mid S \iff \theta_{A \mid S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \mid S} \mid \theta_S.$$

Chordal graphs

If $\mathcal{G}$ is chordal and θ is hyper Markov on $\mathcal{G}$, it holds that

$$A \perp_{\mathcal{G}} B \mid S \Rightarrow A \perp\!\!\!\perp_{\mathcal{L}} B \mid S$$

i.e. it is not necessary to specify that S is a complete separator to obtain the relevant conditional independence.

This follows essentially because for a chordal graph it holds that

$$A \perp_{\mathcal{G}} B \mid S \Rightarrow \exists S^* \subseteq S : A \perp_{\mathcal{G}} B \mid S^* \text{ with } S^* \text{ complete.}$$

If $\mathcal{G}$ is not chordal, we can form $\overline{\mathcal{G}}$ by completing all prime components of $\mathcal{G}$.

If θ is hyper Markov on $\mathcal{G}$ it is also hyper Markov on $\overline{\mathcal{G}}$, and thus

$$A \perp_{\overline{\mathcal{G}}} B \mid S \Rightarrow A \perp\!\!\!\perp_{\mathcal{L}} B \mid S.$$

But the similar result is *false* for an arbitrary chordal cover of $\mathcal{G}$.

Directed hyper Markov property

We have similar notions and results in the directed case.
Say $\mathcal{L} = \mathcal{L}(\theta)$ is *directed hyper Markov* w.r.t. a DAG $\mathcal{D}$ if θ is directed Markov on $\mathcal{D}$ for all $\theta \in \Theta$ and

$$\theta_{v \cup \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{nd}(v)} \mid \theta_{\mathrm{pa}(v)},$$

or equivalently $\theta_{v \mid \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{nd}(v)} \mid \theta_{\mathrm{pa}(v)}$, or equivalently for a well-ordering

$$\theta_{v \cup \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{pr}(v)} \mid \theta_{\mathrm{pa}(v)}.$$

In general there is no similar statement corresponding to the global property and d-separation.
However, *if $\mathcal{D}$ is perfect, $\mathcal{L}$ is directed hyper Markov w.r.t. $\mathcal{D}$ if and only if $\mathcal{L}$ is hyper Markov w.r.t. $\mathcal{G} = \sigma(\mathcal{D}) = \mathcal{D}^m$.*

The distributions of maximum likelihood estimators are important examples of hyper Markov laws. But for *chordal graphs* there is a canonical construction of such laws.

Let $\mathcal{C}$ be the cliques of a chordal graph $\mathcal{G}$ and let $\mathcal{L}_C, C \in \mathcal{C}$ be a family of laws over $\Theta_C \subseteq \mathbb{P}(\mathcal{X}_C)$.

The family of laws are *hyperconsistent* if for any C and D with $C \cap D = S \neq \emptyset$, $\mathcal{L}_C$ and $\mathcal{L}_D$ induce the same law for θ_S.

If $\mathcal{L}_C, C \in \mathcal{C}$ are hyperconsistent, there is a unique hyper Markov law $\mathcal{L}$ over $\mathcal{G}$ with $\mathcal{L}(\theta_C) = \mathcal{L}_C, C \in \mathcal{C}$.

In some cases it is of interest to consider a stronger version of the hyper Markov property.

A hyper Markov law is *strongly hyper Markov* if $\theta_{A \mid S} \perp\!\!\!\perp_{\mathcal{L}} \theta_S$ for all complete separators S.

A directed hyper Markov law is *strongly directed hyper Markov* if $\theta_{v \mid \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{pa}(v)}$ for all $v \in V$.

Parameter $\theta \in \Theta$, data $X = x$, likelihood

$$L(\theta \mid x) \propto p(x \mid \theta) = \frac{dP_\theta(x)}{d\mu(x)}.$$

Express knowledge about θ through a *prior law* π on θ. Use also π to denote density of the prior law w.r.t. some measure ν on Θ. Inference about θ from x is then represented through *posterior law* $\pi^*(\theta) = p(\theta \mid x)$. Then, from Bayes' formula

$$\pi^*(\theta) = p(x \mid \theta)\pi(\theta)/p(x) \propto L(\theta \mid x)\pi(\theta)$$

so the *likelihood function is equal to the density of the posterior w.r.t. the prior* modulo a constant.

A family $\mathcal{P}$ of laws on Θ is said to be *stable under sampling* from x if

$$\pi \in \mathcal{P} \Rightarrow \pi^* \in \mathcal{P}.$$

If the family of priors is parametrised:

$$\mathcal{P} = \{P_\alpha, \alpha \in \mathcal{A}\}$$

we sometimes say that α is a *hyperparameter.* Then, Bayesian inference can be made by just updating hyperparameters.

If $\mathcal{L}$ is a prior law over Θ and $X = x$ is an observation from θ, $\mathcal{L}^* = \mathcal{L}(\theta \,|\, X = x)$ denotes the *posterior law* over Θ.

If $\mathcal{L}$ is hyper Markov w.r.t. $\mathcal{G}$ so is $\mathcal{L}^*$.

If $\mathcal{L}$ is strongly hyper Markov w.r.t. $\mathcal{G}$ so is $\mathcal{L}^*$.

In the latter case, *the update of $\mathcal{L}$ is local to prime components*, i.e.

$$\mathcal{L}^*(\theta_Q) = \mathcal{L}^*_Q(\theta_Q) = \mathcal{L}_Q(\theta_Q \,|\, X_Q = x_Q)$$

and *the marginal distribution p of X is globally Markov w.r.t. $\overline{\mathcal{G}}$*, where

$$p(x) = \int_\Theta P(X = x \,|\, \theta)\mathcal{L}(d\theta).$$

For a k-dimensional exponential family

$$p(x \mid \theta) = b(x) e^{\theta^\top t(x) - \psi(\theta)}$$

the *standard conjugate family* is given as

$$\pi(\theta \mid a, \kappa) \propto e^{\theta^\top a - \kappa \psi(\theta)}$$

for $(a, \kappa) \in \mathcal{A} \subseteq \mathcal{R}^k \times \mathcal{R}_+$, where $\mathcal{A}$ is determined so that the normalisation constant is finite.

The conjugate family is stable under sampling and posterior updating from $(x_1, \ldots, x_n)$ with $t = \sum_i t(x_i)$ is then made as $(a^*, \kappa^*) = (a + t, \kappa + n)$.

Hyper inverse Wishart and laws

Gaussian graphical models are canonical exponential families. The standard family of conjugate priors have densities

$$\pi(K \mid \Phi, \delta) \propto (\det K)^{\delta/2} e^{-\operatorname{tr}(K\Phi)}, K \in \mathcal{S}^+(\mathcal{G}).$$

These laws are termed *hyper inverse Wishart laws* as Σ follows an inverse Wishart law for complete graphs.

For chordal graphs, each marginal law $\mathcal{L}_C$ of Σ_C is inverse Wishart.